REPORT RESUMES

ED 003 279 24
COMPARATIVE EFFECTIVENESS OF COMPOSITION SKILLS LEARNING
ACTIVITIES IN THE SECONDARY SCHOOL.
BY- MCCOLLY, WILLIAM REMSTAD, ROBERT
WISCONSIN UNIV., MADISON
REPORT NUMBER CRP-1528 PUB DATE 63
EDRS PRICE MF-$0.18 HC-$4.20 105P.

DESCRIPTORS- *COMPOSITION SKILLS (LITERARY), *ENGLISH INSTRUCTION, *HIGH SCHOOLS, CONVENTIONAL INSTRUCTION, LEARNING, *TUTORING, *WRITING EXERCISES, WRITING SKILLS, SECONDARY SCHOOL STUDENTS, MADISON, WISCONSIN

CONTROLLED CLASSROOM EXPERIMENTS ATTEMPTED TO ESTABLISH A BASIS FOR EVALUATING THE COMPARATIVE EFFECTIVENESS OF THE WRITING ACT ITSELF, DRILL ON USAGE AND MECHANICS, SELF-INSTRUCTION, GROUP DISCUSSION, THEME CORRECTION AND CRITICISM, AND IMMEDIATE TUTORIAL FEEDBACK. EXPERIMENTS WERE CONDUCTED WITH ALMOST 300 STUDENTS IN GRADES 8 THROUGH 12. IT WAS FOUND THAT (1) ADDITIONAL WRITING ALONE DID NOT IMPROVE WRITING, (2) TUTORING, EVEN WHERE IMMEDIATE FEEDBACK WAS POSSIBLE, WAS OF MARGINAL VALUE AT BEST, AND (3) FOUR FUNCTIONAL COMPOSITION LEARNING-ACTIVITIES CONSISTING OF PRACTICAL STUDY, GROUP DISCUSSION, SELF-INSTRUCTION, AND CONVENTIONAL TEACHER CORRECTION PROVIDED AN INCREASE IN WRITING ABILITY. (JM)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

COMPARATIVE EFFECTIVENESS OF COMPOSITION SKILLS LEARNING ACTIVITIES IN THE SECONDARY SCHOOL

Cooperative Research Project 1528

William McColly
Project Director and Chief Investigator

and

Robert Remstad
Project Statistician

University of Wisconsin
Madison

1963

The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare.

ED003279

PREFACE

This is a report of research conducted at the University of Wisconsin from the summer of 1961 to the fall of 1963. On August 10, 1961, the University submitted to the Commissioner of Education the proposal that this project be supported by federal funds through the Cooperative Research Program. The proposal was approved in December, 1961, and was revised on January 9, 1962, by Dr. William McColly, the project director and chief investigator, and Dr. William Carriker, the representative of the Cooperative Research Branch, to become an appendix to the contract between the University and the Office of Education.

The core of this research was classroom experimentation involving nearly all students at the University of Wisconsin High School, Madison, for the entire academic year 1962-63. The planning stage was formally begun on February 1, 1962, and the analysis and interpretation of data were finished in October, 1963.

Many people besides the authors have spent considerable time on this study: Professor Julian Stanley, as design and analysis consultant; Professor Philip Lambert, as administrative

consultant; Professor Chester Harris, as Director of the Statistical Laboratory of the School of Education; Professor John Goldgruber, as Principal of the Wisconsin High School; Professor Eric Kurtz, of the Department of English of Mount Holyoke College, as a reading supervisor; Mr. David Wiley, as design specialist; Miss Betty Irwin, Mr. Paul Rathburn, and Mr. James Richards, as research and administrative assistants; and Mrs. Elizabeth Dowling, Mrs. Marjorie Elliott, Miss Mary Henley, Mrs. Janet Kurtz, Mr. Theodore Lemcke, and Mrs. Alice Mielcarek, the Wisconsin High School English teachers.

A special debt of gratitude is owed the design and analysis consultant, Professor Stanley. Whatever contribution this research may make to the field of education will have derived substantially from his knowledge and inspiration. Also, many thanks are due the following scholars for their valuable suggestions during the writing of this report: Dr. Robert Beezer, of the Cooperative Research Branch; Professor Carl Bereiter, Department of Educational Psychology, University of Illinois; Professor Eric Klinger, Department of Psychology, University of Minnesota at Morris; Professor Ellis Page, Director of the Bureau of Educational Research, University of Connecticut; and Professor Dallas Wiebe, Department of English, University of Cincinnati. Finally, the project director

should like to particularly thank Professor Robert Pooley and Dean Lindley J. Stiles for their generous advice and support.

TABLE OF CONTENTS

Preface ii
List of Tables vii
PART ONE: THE PROBLEM UNDER FOCUS 1
PART TWO: OBJECTIVES OF THE STUDY 4

PART THREE: RELATED RESEARCH

A. Review and Discussion of Previous Research 6
B. Conclusions Based on Previous Research Studies 16
C. Relationships Between Conclusions Drawn from Previous Research and the General Design of the Study 16

PART FOUR: PROCEDURES

A. Planning 21
B. The Classroom Experiments 28
C. Gathering Post-Experimental Data for Analyses 39
PART FIVE: ANALYSES OF THE DATA AND FINDINGS 49
PART SIX: CONCLUSIONS AND IMPLICATIONS 62

PART SEVEN: APPENDICES

A. Tables Not Included in Text, Experiment One 67
B. Tables Not Included in Text, Experiment Two 72

C. Pilot Experiment to Compare 4-Point and 6-Point Composition Rating Scale 85

D. Impromptu Theme Topics and Writing Situations 88

E. Summary Description of Population Involved in Experiment 90

F. Wisconsin High School English Teachers Participating in Experiment 91

G. Composition Reading Teams 92

BIBLIOGRAPHY 93

LIST OF TABLES

1. Design for Reading of All Antecedent and Criterion Themes 43

2. Numbers of Subjects in Factor-Level Combinations for 8th and 9th Grade Experiment 51

3. Analysis of Variance Results for Experiment One Using the Mean Score on Two Criterion Themes as the Dependent Variable 52

4. Analysis of Covariance Results for Experiment One Using Mean Scores on Two Criterion Themes as the Dependent Variable 54

5. Numbers of Subjects in Factor-Level Combinations for Experiment in Grades 10, 11 and 12 55

6. Analysis of Covariance Results for Experiment Two Using Mean Scores on Two Criterion Themes as the Dependent Variable 57

7. Analysis of Variance Results Using ITED #7, Ability to Interpret Literary Materials, as the Dependent Variable (Grades 8 and 9) 58

8. Analysis of Covariance Results Using ITED #7, Ability to Interpret Literary Materials, as the Dependent Variable (Grades 10, 11 and 12) 59

9. Experiment One. Numbers of Subjects in Each of the Factor-Level Combinations for the 8th and 9th Grade Experiment 67

10. Experiment One. Mean General Merit Scores on Criterion Themes for All Factor-Level Combinations 68

11. Experiment One. Mean General Merit Scores on Antecedent Theme A for All Factor-Level Combinations 68

12. Experiment One. Mean General Merit Scores on Antecedent Theme B for All Factor-Level Combinations 69

13. Experiment One. Mean Scores on SCAT Verbal Subtest for All Factor-Level Combinations 69

14. Experiment One. Mean Scores on Nelson-Denny Reading Test for All Factor-Level Combinations 70

15. Experiment One. Mean Scores on ITED #3, Correctness and Appropriateness of Expression, for All Factor-Level Combinations 70

16. Experiment One. Mean Scores on ITED #7, Ability to Interpret Literary Materials (June, 1963) 71

17. Experiment Two. Numbers of Subjects in Each of the Factor-Level Combinations 72

18. Experiment Two. Mean General Merit Scores Given Criterion Themes by Four Reader Teams 73

19. Experiment Two. Mean General Merit Scores Given Antecedent Themes by Four Reader Teams 74

20. Experiment Two. Mean Scores on ITED #3, Correctness and Appropriateness of Expression, for All Factor-Level Combinations 75

21. Experiment Two. Mean Scores on Nelson-Denny Reading Test for All Factor-Level Combinations 76

22. Experiment Two. Mean Scores on SCAT Verbal Subtest for All Factor-Level Combinations 77

23. Experiment Two. Mean Scores on SCAT Quantitative Subtest for All Factor-Level Combinations 78

24. Experiment Two. Mean English Grade-Point Average for All Factor-Level Combinations 79

25. Experiment Two. Mean Scores on ITED #7, Ability to Interpret Literary Materials (Sept., 1962) 80

26. Experiment Two. Mean Scores on ITED #7, Ability to Interpret Literary Materials, (June, 1963) 81

27. Experiment Two. Mean Number of Misspellings per Criterion Theme for All Factor-Level Combinations 82

28. Experiment Two. Mean Number of Wrong Word Forms per Criterion Theme for All Factor-Level Combinations 83

29. Experiment Two. Mean Number of Sentence Errors per Criterion Theme for All Factor-Level Combinations 84

I. THE PROBLEM UNDER FOCUS

To teach written composition in the secondary school requires the selecting and managing of intricate combinations of learning activities. The teacher and student activities that can be included in the teaching of composition may be classed as follows:

1) Writing per se (writing activity alone, or "mere writing")
2) Study of grammatical, rhetorical, and stylistic theory as theory
3) Practical study of grammar, rhetoric and style: teacher explanation of and student practice on points and matters of usage, mechanics, organization, sentence structure, etc.
4) Group discussion
5) Self-instruction: correcting, editing, revising, and rewriting
6) Conventional teacher correction, criticism and evaluation of writing (oral or in theme annotation)
7) Tutoring with immediate feedback: individualized instruction on a 1:1 teacher-student basis, the teacher providing momentary reinforcement while the student is attempting to compose

When a high school English teacher tries to determine the relative functions and to predict the appropriate amounts of these composition skills learning activities, he is handicapped by a lack of knowledge. Because few experiments on composition teaching have been conducted during the recent period of development in research methodology, teachers must largely use ideas based not on systematic research but on faith, intuition and personal experience. The teaching of composition suffers from a critical need for reliable information about the value of different kinds of learning activities.

Of special necessity is decisive knowledge about the writing act, about the value of writing in and of itself, isolated from other activities. This need has arisen from the current movement to upgrade the teaching of high school English. There is now a popular assumption, held by many educators and English teachers, and found often in the professional literature, that the mere act of writing results in superior development of writing skills.[1] Pressed to assign themes only for the sake of enforcing constant writing, teachers should be given the knowledge to help them judge confidently whether, as is claimed, students who simply write more will become better writers than those who write less, or whether the mere act of writing has no benefit as a composition skills learning activity.

---

[1] For example, according to the "Dusel Report," a widely cited survey of "430 experienced teachers and nationally known teacher-educators concerning the approximate amounts of writing practice considered necessary . . . . Many recommend that some writing be done daily." William Dusel, "Determining an Efficient Teaching Load in English," _Ill_. _English_ _Bull_., Oct., 1955, p. 4.

In brief, the research reported here was focused on a general problem whose solution is basic to building a sound composition program and to improving composition instruction in the schools: of the kinds of learning activities that may be used in teaching composition, which ones are effective and which ones are not?

II. OBJECTIVES OF THE STUDY

Because the problem under focus is so complex and has been the subject of so little research, this project was conceived as a pilot study, as a necessary step before more costly and sensitive experiments can or should be undertaken.

The main purpose of the study was to establish, through controlled classroom experiments, a sound basis for evaluating the effectiveness of composition skills learning activities. The general objectives were 1) to compare the effects of significantly different amounts of different kinds of activities at different grade levels, and 2) to test the popular assumption that the writing act alone is a beneficial learning activity. In other words, this research attempted to answer these questions:

1) What are the comparative effects of certain kinds of composition skills learning activities as inferred from controlled variations of the amount and frequency of these activities at more than one grade level?

2) Which kinds of composition skills learning activities can and should be further investigated in a long-range, large-scale classroom experiment in public schools?

As the project got under way, it became apparent that it might contribute ideas and insights of value in identifying problems of design and analysis in this kind of experiment. This contribution has been kept in mind, then, as a secondary purpose of the study. Certain procedures are described in more than ordinary detail because of this aim.

III. RELATED RESEARCH

A. Review and Discussion of Previous Research

1. Studies of Effects on Composition Skills Learning of Different Amounts and Frequencies of Writing

Earlier research warrants the following inferences. At the college level, a substantial increase in writing activity alone has no apparent beneficial outcomes. At the high school level this may be true also, or it may be that for most students the value of increasing their amounts and frequencies of writing per se diminishes to an imperceptible point at some time during the high school years.[2]

As for research on college students, the earliest significant study is a 1950 survey of the 2400-odd freshmen at Michigan State University.[3] Questionnaires asked each subject for the amount of time he had spent during the academic year on expository writing aside from time spent on it in the freshman communications course. The authors then selected two groups, the 99 subjects reporting the most writing and the 99 subjects reporting the least. The former group reported a range in hours of 71 to 301 with a mean of 131. The latter group reported a range of 0 to 13 with a mean of 4. The authors found that the two groups had done equally well in the freshman communications course and had a mean general grade point average for the year whose difference is insignificnat.

---

[2] In an article appearing after the present study had begun, Frank Heys reports that a small experiment in which significantly different writing amounts were given two 11th grade groups had negative results ("The Theme a Week Assumption," English Journal, May, 1962, pp. 320-22).

[3] Paul Dressel and others, "Effect of Writing Frequency upon Essay-Type Writing Proficiency at the College Level," Journal of Educational Research, December, 1952, pp. 285-93.

The author of an unpublished dissertation investigated the high school English curriculums of the University of Missouri freshmen for the years 1950-51 and 1951-52 with regard to their scores on University placement and matriculation tests in English.[4] The author grouped the high school curriculums according to amounts of writing, grammar and literature, and size of school. He grouped the test scores from rank order. He found no significant correlation between the test scores and the amount of writing done by the students in high school, but found a significant positive correlation between high test scores and the amount of literature studied in high school.

During the year 1956-57, the author of an unpublished dissertation investigated the effects of varying writing frequency and the amounts of certain other composition skills learning activities in a controlled experiment with 257 freshmen in the University of Alberta.[5] The author randomly divided subjects into three groups of 86, 86, and 85 each. All groups took the regular English curriculum for the semester. The control group did no extra writing. The two experimental groups wrote one extra essay of about 500 words during each of the 16 weeks, on topics assigned by instructors in another academic course. These extra essays were read by the assistants of the non-English instructors. The essays of experimental group one were returned with a letter grade and a critical paragraph from the reader. The essays of experimental group two were returned with a letter grade and a critical paragraph, but in addition every error had been marked and the students in this group spent from 30 to 50 minutes of class time in discussing and revising each essay. On a final criterion essay, the control and first experimental groups apparently showed little difference. The second experimental group scored significantly higher than the other two on some of the factors evaluated.

---

[4] Ned Schrom, Missouri High School English Offerings Reported by University of Missouri Freshmen in Relation to Their Freshman English Placement Test Scores, University of Missouri, 1953.

The authors who report an experiment conducted in freshmen English classes at Purdue University in 1948 claim that

> With the method of double-writing, student failures can be reduced 66 percent, and student improvement, judged by the final semester grade, can be increased 60 percent.[6]

---

[5] Earl Buxton, An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Students' Skill in Written Expression, Stanford University, 1958.

[6] Virgil Lokke and George Wykoff, "Doubling Writing in Freshman Composition--an Experiment," School and Society, December 18, 1948, pp. 437-39.

In this study, the 20 students in the experimental class wrote from 32 to 34 themes during the semester. Matched with each of these subjects were control subjects taken from among seven classes that wrote the regular number of themes, 16. The authors present the following data:

Gradling scale	Number of experimental subjects receiving each grade at midterm	Number of experimental subjects receiving each grade at end of term	Number of control subjects with each final grade
5 (good)	0	2	1
4 (above ave.)	4	5	10
3 (below ave.)	4	9	17
2 (poor)	9	2	6
1 (bad)	3	2	6
	20	20	40

Thus, the percentage 66 represents the difference between the number of experimental students with grades of 1 or 2 at midterm and the number who finally failed. The percentage 60 represents the number of experimentals who raised their grades.

When these data are re-analyzed with the grades of control subjects included, what they really show is that the difference between the control and experimental groups' grades can be expected to occur 9 out of 10 times simply by chance. (A chi-square computed from the distribution of final grades for both groups is equal to .7692 with four degrees of freedom.)

Other negative factors might be considered, also. Because the authors used as criteria the grades given by the experimental class' teacher, there was possibly a "lesson-learning" effect (i. e., from learning to do what a teacher specifically wants). This kind of evaluation may have created a halo effect (from a teacher's feeling that his efforts inevitably get results). Because the control subjects, apparently, were naive while the experimentals knew they were involved, there was a possible Hawthorne effect. Finally, there may have been a matching error factor.

All in all, when these things are considered along with the chance factor, for what it is worth the Lokke and Wykoff experiment can be interpreted as really confirming the null hypothesis about doubling writing for a semester. Their study has been analyzed in detail here because their conclusions have been accepted without question in certain influential publications.

2. Studies of the Effects of Theoretical Language Study on Composition Skills Learning

It seems fairly well established that the relationship between a student's knowledge of language theory and his ability to write is not causal, as many people assume, but really is what might be termed "cognate." That is, in the order of mental abilities, ability to learn about language and ability to use language are seen as having much the same rank and as stemming from some more primary factor. The still popular assumption that students have to know "grammar" in order to learn to write is untenable. For many years, research on the hypothesized causal relationship between theoretical study and writing ability has yielded generally negative results. Reports and discussions of this research can be found in Rollo Lyman (1929, 1931), Dora Smith (1938), Robert Pooley (1957), and Ingrid Strom (1960).[7]

3. Studies of Effects of Practical Instruction (as Distinguished from Instruction in Theory) on Composition Skills Learning

The learning activities of this kind are teacher explanation of and student practice on points and matters directly relevant to written expression, from things as finite as spelling to those as broad as paragraph organization. The basic activity in this kind of instruction is student practice, with the teacher's role being to manage the practice and to provide the student with knowledge necessary for making the practice as meaningful as possible. By the time of Lyman's summary, evidence suggested that this sort of activities results in learnings that can be put to use in composition. This conclusion is reached also by Smith and Strom. Pooley's Teaching English Grammar is a treatment of the practical approach to language study as distinguished from the theoretical. It might be noted here that the recently developed foreign language learning programs based on structural linguistics, including those for the teaching of English as a foreign language, exclude the study of theory as theory, but rely on "pattern practice" or "structure drill" as a basic activity.

---

[7] Lyman, Summary of Investigations Relating to Grammar, Language and Composition. University of Chicago Press, 1929 (Supplement, 1931).
Smith, "English Grammar Again!," English Journal, October, 1938, pp. 643-49.
Pooley, Teaching English Grammar. Appleton-Century-Crofts, 1957.
Strom, "Research in Grammar and Usage and Its Implications for Teaching Writing," Indiana University School of Education Bulletin, 1960.

4. Studies of Effects on Composition Skills Learning of Group Discussion

A small amount of research summarized by Lyman, by Strom and by others, and studies such as the Buxton experiment (see pages 7 and 8) and the Maize experiment (reviewed on the following page) indicate value in this kind of activity. The exact role of discussion in teaching composition has not been determined, however. It may not be a composition skills learning activity as such, but one that merely expedites this learning. If the latter is true, discussion may have great value as a means of presenting information, as a motivating force, and as an adaptive mechanism facilitating the content of composition.

5. Studies of Effects on Composition Skills Learning of Self-Instructional Activities

That the student activities of correcting editing, revising, proof-reading, and rewriting should have positive operant effects might be assumed, and research summarized by Lyman, by Smith and by Strom tends to confirm this. The value of these activities is implied also in results of an experiment conducted in remedial freshman English classes at Purdue University during the year 1949-50. To a control group of 74 students the author gave a curriculum whose emphasis was the study of theory. To a matched group of 75 he gave a curriculum comprised mainly of self-instructional activities, conference and discussion. Terminal compositions written by the experimental students are reported to be significantly better than those written by the control students.[8]

6. Studies of Effects on Composition Skills Learning of Conventional Teacher Feedback through Theme Correction, Criticism and Evaluation

These activities, whose description might be shortened to "theme correction," promise to be beneficial according to results of studies summarized by Lyman and by Strom. This kind of effect can be inferred, also, from research done outside the field of English composition. In a large classroom experiment, Ellis Page sought answers to these questions:

> 1. Do teacher comments cause a significant improvement in student performance? 2. If comments have an effect, which comments have more than others . . . ?

Each teacher involved agreed to treat the examination papers in one of his classes in the following way. To 1/3 of the papers for the class selected the teacher assigned randomly, from casting a special die, 1) a grade but no comment, 2) to another third a grade plus whatever comment the teacher wanted to write, and 3) to the other third a grade plus a one-phrase comment already made up for that particular grade by the experimenter. The outcome was that on the examinations next given in those particular classrooms the students who had received comments scored significantly higher than the students who had received a grade alone. There were no significant measurable differences resulting from the two kinds of comments, free or formalized (i. e., genuine or gratuitous).[9]

---

[8] Ray Maize, "Writing Laboratory for Retarded Students," College English, October, 1954, pp. 44-48.

Robert Bostrom and others investigated the possible reinforcement effects of different grades. They gave university students an attitude questionnaire, then had the students write essays defending positions contrary to those indicated by the survey. They reported either a good or a poor grade randomly to each student. From a second survey, they found that the students receiving good grades had changed attitude much more than had those receiving poor grades.[10]

---

[9] "Teacher Comments and Student Performance: a Seventy-four Classroom Experiment in School Motivation," Journal of Educational Psychology, August, 1958, pp. 173-81.

[10] "Grades as Reinforcing Contingencies and Attitude Change," Journal of Educational Psychology, April, 1961, pp. 112-15.

7. Studies of Effects of Immediate-Feedback Tutoring on Composition Skills Learning

In the teachingof composition, tutoring in which "immediate" feedback is really possible must take place during the time a student is actually composing. This is to say, while a student is writing or attempting to write the teacher corrects an error just made, praises a clearly expressed thought just written down, etc. This kind of tutoring seems quite uncommon in the teaching of writing. In fact, no research studies really identify the activity. Most tutoring is done with comparatively long feedback delay. At best, a teacher and student confer about something written during the day of a conference, but often they are forced to take up writing that has lain about for a time. In the present study, the customary kind of tutoring is classed simply as an oral variety either of theme correction or of practical instruction.

The basis for regarding immediate-feedback tutoring as a kind of learning activity distinct from the conventional mode of tutoring is the theory that feedback effect depends to some extent on avoidance of delay. It is reasonable to assume that a student will be decidedly more influenced by information about his errors and progress given as he composes than by feedback he gets after he has finished a composition and put it aside.

B. Conclusions Based on Previous Research Studies

From the foregoing review and discussion, the following inferences can be drawn about the effects of various kinds of activities included in teaching composition skills in the secondary school:

1) No measurable benefit is derived from the study of theory as theory.

2) The value of "mere writing" is doubtful.

3) Some value results from these four kinds of activities:

   Practical instruction

   Group discussion

   Self-instruction

   Teacher theme-correction of the conventional type

4) Immediate-feedback tutoring, in which momentary reinforcement is given during composition, is perhaps the most beneficial kind of composition skills learning activity.

C. Relationships Between the Conclusions Drawn from Previous Research and the General Design of This Study

On the basis of the first conclusion, the study of theory per se was rejected as a possible experimental variable. It appeared necessary and profitable to measure the effects of different amounts of the other six kinds of activities only.

Because many teachers have great faith in the value of writing activity alone and emphasize mere writing in teaching composition, and because the negative results of research on college subjects implies that this activity lacks value also in the high school, it was thought appropriate to design the experiment so that writing per se would be the single experimental variable at two or more grade levels.

Of the other five kinds of activities, which might be called "non-writing" activities,[11] immediate-feedback tutoring stands apart. First, no research literature deals with it (as a composition skills learning activity, that is). Second, reinforcement theory suggests that it might prove more effective than most composition learning activities. Third, its great cost makes it the least feasible activity of all. Because it seemed important to establish an experimental basis for predicting the value of this kind of tutoring, it was decided to treat it as an experimental variable at several grade levels.

The remaining four activities form a distinct class. First, they stand out as the activities indicated by previous research as having value in composition instruction. Second, they are closely related and can make a well-integrated sequence of activities. Once their management is demonstrated to a teacher, he can then unite them through a composition task. He assigns or elicits a writing task during which students apply points recently explained and practiced on, from which matters arise for discussion, whose writing in progress is the vehicle for revising, editing, etc., and whose finished product serves as a basis for criticism and rewriting. Thus, these four activities can be patterned for a common purpose, which is to affect the function of writing ability in a meaningful situation. For this reason, they should be termed the "functional" composition learning activities. (Immediate-feedback tutoring is in this sense functional, too.) The approach to composition teaching in which these four activities are managed through a writing task should be called the functional method.

---

[11] They are "non-writing" in the sense that assuming the distinctiveness of writing per se and recognizing its isolation as an experimental variable make them so.

From the facts that these activities can be managed as a closely related group and the present study is pilot research with the role of breaking ground for more complicated experiments, they were not considered as four separate activities but altogether, as one experimental variable to be treated at several grade levels.

From the various kinds of composition skills learning activities, then, three experimental variables were chosen, and experiments were designed to test the following hypotheses:

1) More writing alone results in better writing.

2) More functional "non-writing" composition learning activities result in better writing.

3) Tutoring with immediate feedback has significant benefit as a composition skills learning activity.

It was decided that certain English classes at six grade levels, 7 through 12, at the University of Wisconsin High School, Madison, should be used for the entire academic year 1962-63, according to the following general plan. To test the hypothesis that more writing means better writing, a pair of classes at each of two or three grade levels would be given identical curriculums except for the amounts and frequencies of writing per se. The control class in each of these pairs would write about 1/4 as much as the experimentals. Control students would be assigned a composition of 300 words or less about once a month, and the experimentals one of 300 words or more about once a week. These amounts and frequencies were felt to be most appropriate because of the "theme a week" idea--the common conviction that weekly writing is imperative, while writing as seldom as once a month may really be fruitless--and because of the general use of the 300-word assignment.

With each of the three or four grade levels remaining, plans were made to use a set of three classes for experiments on the "non-writing" activities. Each set of three would have one control class and two experimental classes. The amount and frequency of writing in this experiment would also be given in the approximate 1:4 ratio, but the compositions would serve merely as an operational basis for managing different amounts of the group of four functional activities on the one hand, and of tutoring on the other. In view of the recommendation in Conant's The American High School Today (1959) that 1/2 the time given to the subject of English be spent on teaching composition, it seemed appropriate to establish for the sets of classes in this experiment the following amounts of the four functional activities. The experimantal classes would write about a theme a week, this task to be the vehicle for about 2 and 1/2 days of the functional activities. The control classes would have a monthly task of the same kind. Actually, they would undertake only every fourth weekly task, and thus have about 1/8 of their time spent in functional composition instruction.

To measure the effects of immediate-feedback tutoring, plans were made to tutor all the students in one of the two experimental classes in each of these sets, and to withhold activity of this kind from the control classes and other experimental classes. The students to be tutored would meet with their teacher in groups of three for one class period (56 minutes), when the teacher would try to spend 1/3 of the period with each student on an individual basis, sitting by him to give immediate reinforcement during his writing.

This general design was felt to be inherently productive of data from which tests of the three hypotheses could be made and analyses done of the effects of different amounts of different activities at different grade levels. It became the framework of research undertaken through the following procedures.

IV. PROCEDURES

A. Planning

From February through May, 1962, the principal investigator and other project staff reviewed literature dealing with the teaching of language and composition. One aim of this was to ensure that nothing substantial had been overlooked in conceiving the rationale and general design of the experiment. Another aim was to help enlarge our fund of specific classroom activities that English teachers find manageable and that show promise of being effective.

The main result of reviewing a great number of items--books, articles, monographs and unpublished materials--was a confirmation of the chief investigator's earlier impression that literature on the teaching of English is typically vague and general. No research studies were uncovered that suggested modifying the design or hypothetical base. Some practical items reporting techniques favored by teachers were abstracted for use in planning detailed activities for the classrooms.

While searching this professional literature, the chief investigator began to formulate curriculum plans for control and experimental classes. The aim of this was to ensure proper managing of the experimental variables during the classroom experiments. Certain curricular procedures and specific classroom activities that resulted from these plans are described at appropriate points in later sections of this report.

In February, 1962, the chief investigator and the design and analysis consultant began to think through the task of choosing subjects and assigning them to control and experimental classes. The original plans (as stated in the contract) called for using students in all six grades of the Wisconsin High School,[12] involving two or three classes at each grade level, giving a total number of subjects of about 300. The method of assigning these students to control and experimental classes was to be the following. Students were to be matched individually, then randomly assigned through a randomized block design. The names of all students in each grade were to be listed in rank order from the totals of average standard scores of two tests, one for English composition skills and one for general vocabulary. For assigning to a two-class set, student 1 was to be a control, student 2 an experimental, student 3 a control, student 4 an experimental, etc. For assigning to a three-class set, the same kind of procedure was to be used. This testing and assigning was to be done in May and June of 1962.

---

[12] A general description of this population and of the High School is in an appendix.

During the working out of details, it became apparent that the randomized block design for assigning subjects was unfeasible, mainly because of the conflict, in the upper grades, between the students' preference or necessity for taking certain courses and the random assignments. In March, 1962, Professor Stanley decided to abandon this plan and to substitute for it the following. Assignments would be made to control and experimental classes as randomly as possible. With this somewhat random grouping, analysis of covariance techniques would be used to remove bias and to increase the precision of the analysis.[13] The particular measures used as covariates are outlined in later sections of this report.

Before the details of this somewhat random grouping are described, a second major departure from the original plans should be explained. About the first of May, 1962, after a University faculty committee had recommended that the Wisconsin High School be dissolved within a reasonable time, the School of Education decided not to admit a new class (of 7th graders) in the fall. This reduced the number of grade levels for 1962-63 to five, 8 through 12. The total population anticipated was slightly more than 300. Because of this situation, the chief investigator and the design and analysis consultant began preparing to use as subjects all students in all English classes at the High School. The experiment to test the hypothesis that more writing alone means better writing was planned for grades 8 and 9, where two classes each of from 20 to 25 students apiece were expected. The experiment to test the other two hypotheses, that more functional instruction results in better writing and that immediate-feedback tutoring is an effective composition learning activity, was planned for grades 10, 11, and 12. Here, three classes each of from 20 to 25 students apiece were expected. The original approximate total number of subjects did not change, but the number of grade levels dropped from six to five.[14]

---

[13] This major departure from the contract appendix plan was stated in the progress report of April 30, 1962.

Accordingly, during the summer of 1962 the 13 English classes scheduled for 1962-63 were made up as randomly as possible through the following procedures. In grades 8 and 9, where nearly all students in each grade took the same courses, random assignments were possible. From the alphabetical rosters for each of these two grades, students were alternately placed by sex to one of the two classes. That is, the first boy went to class one, the second boy to class two, the first girl to class one, the second girl to class two, etc. In grades 10, 11, and 12, this degree of randomness was impossible. The school principal and the project statistician had to assign students to classes in the usual fashion: trial and error, to avoid as many scheduling conflicts as possible. In this process they were rather successful in equalizing the numbers in classes at each grade level, and tried to balance the numbers of each sex. This assigning was not finished until the end of the summer interim.

---

[14] This major departure from the contract appendix plan was stated in the progress report of June 30, 1962.

When classes began at the High School on Monday, September 10, none of the English sections had yet been designated as either a control or an experimental. Again, this was done randomly for grades 8 and 9, and as randomly as possible for 10, 11, and 12. In each of the upper grades, it was necessary to designate as the second experimental, or tutored, group the section whose members had the most study halls coinciding with the teachers' free periods. The other two sections were then randomly designated either control or first experimental.

From the whole assigning process, then, this was the schedule and class make-up for the experiment:

Grade 8

Control class (monthly writers), period 3 (10:00--10:56)	boys	12
	girls	12
Experimental class (weekly writers), period 7 (2:24--3:20)	boys	12
	girls	11

Grade 9

Control class, period 3 (10:00--10:56)	boys	11
	girls	15
Experimental class, period 2 (9:00--9:56)	boys	12
	girls	15

Grade 10

Control class (monthly theme with functional instruction), period 4 (11:00--11:56)	boys	9
	girls	10
First experimental class (weekly theme with functional instruction), period 1 (8:00--8:56)	boys	6
	girls	14
Second experimental class (weekly theme with functional instruction plus tutoring), period 5 (12:24--1:20)	boys	19
	girls	6

Grade 11

Control class, period 1 (8:00--8:56)	boys	14
	girls	6
First experimental class, period 4 (11:00--11:56)	boys	11
	girls	10
Second experimental class, period 5 (12:24--1:20)	boys	9
	girls	10

Grade 12

Control class, period 4 (11:00--11:56)	boys	13
	girls	10
First experimental class, period 2 (9:00--9:56)	boys	11
	girls	13
Second experimental class, period 6 (2:24--3:20)	boys	6
	girls	18

A total of 295 subjects were in the experiment at the outset.

But even after these assignments and designations had been made, the classroom experimentation itself was not initiated until data had been gathered to provide for covariate measures that were used after the experiment in analyzing the effects of the experimental variables. Covariates are antecedent measures, distinct from criterion, or terminal, measures. The practical effect of an analysis of covariance is to equalize various groups for the abilities measured by the covariates. Actually, the computations in covariance analysis are the basis for predicting the performance of individuals after an experiment from their performance on the covariates, or the pre-experimental measures. It seemed desirable to have as antecedent measures, or covariates, scores representing general scholastic aptitude, vocabulary, reading ability, previous achievement in English, previous general academic achievement, and level of writing ability. The following measures of these abilities were taken:

1) School and College Ability Test (forms IA, IIA, and IIIA)
2) Nelson-Denny Reading Test (which contains a vocabulary test)
3) Iowa Test of Educational Development #3, Correctness and Appropriateness of Expression
4) Iowa Test of Educational Development #7, Ability to Interpret Literary Materials
5) Previous English grade-point average, and previous over-all grade-point average
6) Samples of impromtu writing

Between September 10 and 21 all students took the standardized tests.[15] They wrote a period-long impromptu composition on Thursday, September 13, and another on Tuesday, September 18. The instructions given for this writing and the nature of the two topics are the same as those used for two criterion compositions in May, 1963. The details of these four writing situations are in an appendix.

---

[15] It might be noted that although standardized tests were given in May, 1962, as originally planned, the scores were not used because the first plan for assigning subjects was dropped.

One more aspect of the stage preceding the experiment itself needs to be reported. Throughout the planning, the chief investigator put much thought into teacher selection. Competence was always the first criterion, but such factors as tolerance and objectivity got more than ordinary consideration. Because the teacher market in Madison is generally good and the University attracts many qualified teachers seeking part-time work to support graduate study, it was thought feasible to try to hire one teacher per grade level who would have no other teaching assignment. Not simply was this particular aim realized, but a very knowledgeable and cooperative English faculty was selected. Vitae for these teachers are in an appendix.

B. The Classroom Experiments

Actually, two distinct experiments were conducted. One was in grades 10, 11, and 12 on the effects of the four functional composition activities and of tutoring, and the other was in grades 8 and 9 on the effects of writing per se. For the clearest picture of these experiments, they will be described separately.

1. Testing the Hypothesis That More Writing Alone Means Better Writing

The main task facing the 8th and 9th grade teachers was to give an equal amount and the same kind of instruction to the control class and the experimental class, yet have the latter students write four times as much as the former. This meant that each had to equalize between his two groups not only the time spent on non-composition activities but also the time for all composition skills activities except writing per se.

As a means of equalizing everything except writing alone, the teachers used the following methods and procedures.[16] First, the non-composition curriculum each gave his two groups was practically identical. Second, they distinctly marked off sequences of composition skills learning activities not directly associated with particular writing tasks from the sequences that comprised these tasks. In other words, above and beyond the activities they managed in their writing tasks, they gave certain lessons in composition skills. These lessons were rather short, not going beyond one class-period, so that they were started and finished between the ending of one composition task and the beginning of another. The materials in the lessons were of common varieties. Both teachers used text and exercises from the Corbin and Perrin Guide to Modern English. Also, for example, the 8th grade teacher gave a number of lessons through a programed text on punctuation, while the 9th grade teacher gave a dozen or so review lessons on traditional grammar (parts of speech, sentence elements, etc.). Although the teachers could not always give a lesson to both classes on the same day, by the ends of semesters they had given them the same total amounts of composition skills activities lying outside particular writing tasks. Besides helping to equalize non-writing time, another important purpose of this plan was to reassure control students that even though they were writing only one theme a month, they were still getting the fulness of composition instruction their classmates were.

---

[16] It should be noted that what is described here was used by the 9th grade teacher only in the fall semester. Because the second semester of 9th grade English is given over to instruction in speech skills, in the spring the teacher used a different and much simpler method of equalizing time.

As a third technique for equalizing the amounts of non-writing composition activities, these teachers gave writing tasks of a kind that excluded or discouraged the activity of self-instruction (revising, editing, etc.). Many compositions were written in class, and all class writing was impromptu. (On days when experimental students were writing a weekly theme in class, control students were usually allowed free reading during the period.) Compositions written outside class were limited to one-day assignments. Also, probably because the 8th and 9th graders were comparatively naive about composition, the teachers seemed quite successful in keeping students unaware of the great benefit of careful revising.

The fourth strategy these teachers used for equalizing time was rather complicated. They equalized the total time spent by their groups on activities within the writing task sequences. Each writing task was limited to a sequence of four activities: discussion of the topic, the actual writing, theme correction, and terminal discussion. With actual writing as the one experimental variable, done by the controls and experimentals in a weekly:monthly or 1:4 ratio, the teachers had to equalize total times spent in pre-discussion, theme correction, and post-discussion. This required their applying the 1:4 ratio to each theme task as follows. For all take-home composition assignments given the monthly writers, the teachers spent an average of 40 minutes discussing the topic. For all take-home assignments given the weekly writers, they managed an average of 10 minutes in preliminary discussion. This discussion (and necessary instructions) was the only activity that preceded writing, except for activities undertaken independently by the students after class. In correcting monthly compositions, the teachers spent an average of 12 minutes per theme. They spent an average of three minutes per theme correcting weekly writing. For discussing returned themes, they gave an average of 40 minutes to the monthly writing classes, and an average of 10 minutes to the weeklies.

Finally, to make this equalizing of theme-related activities as exact as possible, the teachers prorated the kinds of correction given themes. That is, they chose the theme factors they corrected on the basis of the 1:4

ratio. These factors were divided into four groups: 1) content and organization, 2) syntax and sentence structure, 3) style, usage and vocabulary, and 4) mechanics (including grammar). The teachers corrected monthly themes for all four factors, but corrected weekly themes for only one factor, taking each factor in turn. An aspect of teacher feedback not equalized was letter grades. At the beginning of the experiment the plan was to give grades to all monthly themes, but to just every fourth weekly theme. After a month or so the student reaction was so severe that the teachers began to grade every theme.

The alternative method to equalizing total amounts of composition skills activities other than writing would have been to exclude them from the curriculum. The chief investigator and the design and analysis consultant considered this plan, but rejected it because of the risks involved. Not only would the students have been deprived, but also they might well have vitiated the experiment by becoming hostile and uncooperative.

On Monday, September 24, after the completion of tests and writing samples from which covariates were taken, the 8th and 9th grade teachers began the programs just described. Because their composition tasks were so short and well-defined, they regularized the frequencies of writing except for occasional interruptions from school activities beyond their control, and except for stepping up the frequency at several points to make up for time lost in testing. During the 30 instructional weeks that could be devoted to experimentation (the regular 33 less three for pre- and post-testing), the following number of compositions were written:

	Controls	Experimentals
Grade 8	9	35
Grade 9	8	34

To further ensure maintaining the approximate 1:4 ratio, the original proposal was that the controls would be assigned themes of no more than 300 words, and experimentals themes of at least 300 words. Actually, the controls were given every fourth theme task given the experimentals. It seemed unwise to insist on these rigid word limits on these tasks, to perhaps invite student suspicion and reaction. On these assignments the controls were simply told that they need not write more than 300 words, and the experimentals told that they should write at least 300 words. On other assignments to the experimentals (when the controls were not writing), they were urged to write from 300 to 400 words. Taking into account total words, the controls averaged about 300 per theme and the experimentals somewhat over 300, with the 9th graders averaging more than the 8th graders.

It has been mentioned (footnote 16) that the particular strategies for equalizing total amounts of non-writing composition skills activities used in grade 8 throughout the year were not necessary in grade 9 after the fall semester.

This was because the second semester of 9th grade English was primarily a curriculum in speech skills. To ensure that writing per se was the single variable during the spring semester, the teacher proceeded as follows. She required her experimental students to write out, as a composition, the weekly public-speaking assignment she customarily gave. She required her control class to write out, as a composition, one such assignment per month. She read all the themes she received, but corrected themes only once a month. In other words, she corrected every set of themes from her monthly writers, and one out of four sets of themes from her weekly writers. She returned three out of four sets of themes to her experimental class with no feedback except a letter grade. The latter class accepted the extra writing with no manifest hostility, probably because by then they had grown used to the experiment, because the curriculum had changed, and because the teacher exercised considerable force of authority.

The 8th and 9th grade students had no more writing after Thursday, May 23, except for the two criterion compositions written for the purpose of providing measures of the effects of experimental variables. This terminal phase of the procedures will be described after the experiment with the upper grades has been taken up.

2. The Experiment Conducted in Grades 10, 11, and 12

The experimental variables here were 1) an intact group of four functional composition skills learning activities and 2) immediate-feedback tutoring. The way the former variable was treated will be described first.

As part of the study's planning stage, the chief investigator had worked out a sequential paradigm of the four functional "non-writing" activities within a typical writing task:

Teacher, group or individual activity: conceiving and initiating the task

Teacher activities: analyzing and explaining problems in the task, and instructing as to forms, techniques and materials appropriate or necessary for executing the task

Group activity: discussing the task and the foregoing problems, techniques etc.

Individual activities: conceptualizing (including outlining) and composing, revising, editing, etc.

Teacher activity: theme correcting

Group activity: discussing the finished task

Individual activity: rewriting

During the first two weeks, which were spent on the writing and testing to provide covariates, the chief investigator oriented the teachers to the use of this sequence.

After the experiment itself got under way, the teachers initiated most of their writing tasks on Friday or Monday (for reasons that will be explained when the procedures for tutoring are discussed), and finished them in four or five days. They devoted part or all of the class period immediately following a theme assignment to explaining problems and procedures, giving instruction to be put into practice by the students in their writing, and promoting discussion. On days the first drafts of themes were due, the teachers allowed a certain amount of time for discussion, gave further explanation or instruction, or had students edit and revise. When they returned corrected themes, they again promoted discussion, often on the basis of reading themes aloud. With certain tasks they had an entire class rewrite a particular theme.

The relative times a teacher allotted to the four activities within each task sequence were determined by the teacher from her judgment of what would be most effective. The particular factors (such as paragraphing, or mechanics, or sentence structure) a teacher stressed in any session of practical instruction was also left to her judgment. The theme assignments of all three teachers were predominantly literature-centered. They did allow a few free compositions, but most of the topics chosen by students had to be relevant to some aspect of the literature being studied. To help these teachers give the most promising kind of "non-writing" instruction that made up this experimental variable, and to help them manage the sequences of writing task activities efficiently, the chief investigator provided not only continual advice and consultation but also certain detailed lesson plans.

In order to give the variable of the intact group of four functional activities its proper 1:4 variation between control and experimental classes, the teachers followed this routine. Regardless of the relative times they gave the four kinds of activities within a task, they spent an average of 2 and 1/2 class periods per task, and tried to make each task at least two and no more than three periods in length. Keeping the approximate 1:4 ratio in mind, they assigned every task to their experimental classes, but only one out of four, or about one task per month, to their control classes. Except for unavoidable interruptions in the school schedule, they maintained this frequency rather well. In the spring, to increase the total numbers of themes and amounts of functional activities, they gave the experimental classes two tasks in certain weeks which were devoted entirely to composition instruction. At the end of the 30 instructional weeks in the experiment, the teachers had managed these numbers of sequences of functional "non-writing" composition skills learning activities:

	Controls	Experimentals
Grade 10	9	36
Grade 11	9	36
Grade 12	8	31

The 10th, 11th, and 12th grade teachers treated the other variable, immediate-feedback tutoring, in the following way. To students in their control and first experimental classes, they gave no instruction of this kind. In fact, they avoided scheduling individual conferences with these students altogether. For their second experimental sections, they scheduled sessions in which they met with two or three students (and on rare occasions, to make up time, with four) during a free class period (56 minutes). They did no tutoring within the English class. It was felt that the best time for tutoring during a writing task sequence is after a topic has been established and students have had time to discuss it and think about it, but before they have finished the writing. For this reason, nearly all tutoring sessions were held on Tuesdays, Wednesdays, and Thursdays. The writing tasks were initiated when possible on Fridays or Mondays, with four or five days allowed for completion.

Tutoring began during the last week in September, and lasted through the third week in May. The project administrative assistant kept a log of tutoring time for each of the tutored students. He and the chief investigator often discussed the techniques of this kind of tutoring with the teachers, and occasionally audited tutoring sessions. At the end of the 30 instruction weeks when tutoring was possible, the tutored students had been in an average of 27 or 28 sessions. When normal student absence and the interruptions in a school schedule are considered, the teachers can be credited with maintaining a good pace of regular and make-up sessions. The reality of these sessions was such that in very few of them could the teacher spend exactly the same amount of time with each student. What the teachers tried to equalize was the total of actual tutoring time per student. By keeping track of these individual totals, they were able to motivate the lazy and reticent students to be better prepared for tutoring sessions and to prevent facile and eager students from monopolizing the sessions.

C. Gathering Post-Experimental Data for Analysis

During the spring of 1963, the chief investigator and the design and analysis consultant worked out plans for providing sources of data for analyzing the effects of the experimental variables.

It was felt that the basic data should come from impromptu compositions. Writing done in other kinds of situations promised to be unreliable because of the number of factors (such as outside help) that might affect it. It was decided that the best impromptu writing for criterion purposes would be two themes of substantial length on topics similar to those the students had written on prior to the experiments. These four topics and the writing situations are described in detail in Appendix D.

Accordingly, on Friday, May 24, students in all English classes wrote a period-long impromptu theme on the same topic.

On Monday, May 27, they wrote a second period-long impromptu criterion theme. The number of absentees was low each day, and make-ups were finished by Wednesday, May 29. On both days the teachers told their classes that these themes were part of their final examination for the semester in English.

These criterion themes, along with the two similar themes that had been written in September to provide covariate measures--a total of about 1150--were then typed verbatim, with every error carefully preserved, in standard format.[17] The names were withheld, and each theme was stamped with a randomly chosen identifying number.

Because of the success of the Educational Testing Service in evaluating writing samples, the chief investigator spent a day in Princeton in April to familiarize himself with their operation and to consult with them about the theme reading for this experiment.[18] Making considerable use of their advice and information, Professor Stanley, Mr. Wiley and the chief investigator planned the following design and procedures for rating both the criterion and the covariate impromptu themes.

---

[17] That the typing and standardizing were necessary to remove an important source of bias seems clear. One author reports an experiment in which he factor-analyzed 230 themes rated by 10 readers. The factor appearing most consistently and with the highest loadings was made of these externals: general appearance, clarity of handwriting, and esthetic arrangement. (C. Remondino, "A Factorial Analysis of the Evaluation of Scholastic Compositions in the Mother Tongue," British Journal of Educational Psychology, Nov., 1959, pp. 242-51.

[18] Special thanks go to Mrs. Jean Beck, Miss Frances Swineford, and Mr. Fred Godshalk.

With the aid of Professor Jarvis Bush of Oshkosh, Executive Secretary of the Wisconsin Council of Teachers of English, sixteen theme readers were somewhat randomly chosen from the names of classroom teachers on the rolls of the Council, in these categories: four college males, four college females, four high school males, and four high school females. One person from each category was placed on a team of four readers, four such teams being formed. Each team came to Madison for an evening orientation session, stayed overnight, and read themes the next day. The four reading days were June 26-29. It should be noted here that because the two criterion and the two antecedent themes were typed the same way and were on the same kind of topics, the readers rated them all at the same time. The design for the rating is described in the next few pages.

By the end of the classroom experiments, 283 students remained as experimental subjects out of the original 295. The project statistician made alphabetical lists of the 283 subjects on the basis of grade, sex, and treatment (that is, control and experimental). There were 26 lists: 8th grade boy control, 8th grade girl control, etc. From these lists, he assigned students alternately to four batches. The result of this was four batches (lists of experimental subjects) randomly selected but still having practically equal numbers of subjects from the same grade, of the same sex, and with the same treatment. He then took the four groups of typed themes (all the themes on one topic making a group), and divided each group into four sets according to the names in the four batches of students. The result was 16 sets of themes: those written on the first topic by subjects in batch one, those written on the first topic by subjects in batch two, etc.

The reading was done by four different readers on each of the four days. There were four reading sessions (periods of about 90 minutes) during the day. Thus there were 16 sets of themes, 16 different readers, and 16 different reading sessions. Each set of themes was rated each day. Each set was rated during a session different from the day before. Each set was rated by a type of reader different from the day before (meaning that each theme was rated four times, once by a college male, once by a college female, etc.). For each reading, the order of themes in each set was random. This meant that after each reading a set was re-randomized for the next reading (the order being established by matching the random identifying numbers with a set of numbers from the Rand table of random digits). All in all, no theme was read more than once in the same day, more than once by the same type of reader, more than once during the same period of the day, or, except by chance, more than once in the same relative position in its set. This is a diagram of the reading design:[19]

---

[19] The design is an adaptation of a 2^{10-4} fractional factorial design as outlined by Box and Hunter. (G. E. P. Box and J. S. Hunter, "The 2^{k-p} Fractional Factorial Designs," MRC Technical Summary Report #218, Mathematics Research Center,

Table 1: Design for Reading of All Antecedent and Criterion Themes

		Session (Period of Day)			
Day	Reader	One: about 8:30--10:00	Two: about 10:30--12:00	Three: about 1:15--2:45	Four: about 3:15-4:45
1	1	C4	A3	B2	D1
	2	D4	B3	A2	C1
	3	A4	C3	D2	B1
	4	B4	D3	C2	A1
2	5	D3	B4	A1	C2
	6	C3	A4	B1	D2
	7	B3	D4	C1	A2
	8	A3	C4	D1	B2
3	9	A2	C1	D4	B3
	10	B2	D1	C4	A3
	11	C2	A1	B4	D3
	12	D2	B1	A4	C3
4	13	B1	D2	C3	A4
	14	A1	C2	D3	B4
	15	D1	B2	A3	C4
	16	C1	A2	B3	D4

United States Army, University of Wisconsin, April, 1961.) Following the lead of Box and Hunter that certain interactions could be ignored since it would be reasonable to assume that their effects would be zero or negligible, Wiley drew the following line of reasoning. All second order (i. e., three-factor) and higher interactions were assumed negligible or non-existent. The main effects and first order interactions were characterized as important, not so important, unimportant

The types of readers are:

College male: 1, 5, 9, 13

College female: 2, 6, 10, 14

High school male: 3, 7, 11, 15

High school female: 4, 8, 12, 16

The four groups of themes (by topic) are indicated by the letters A, B, C, and D. The covariate themes written prior to the experiment are A and B. The criterion themes written on May 24 and 27 are C and D. The four batches of experimental subjects, randomly selected but equalized by grade, sex, and treatment, are indicated by the numbers 1, 2, 3, and 4.

The readers evaluated the themes for general merit alone. That is, for each theme they rated they wrote down only one score and that represented their estimate of the over-all writing ability reflected by the theme. The decision to use general merit as a single criterion measure is based for the most part on research done by the Educational Testing Service,

---

on the criterion of expected size of the effect. The design was then arranged so as to free important main effects and first order interactions from as much confounding as possible. All main effects were characterized as important as well as were all first-order (i. e., two-factor) interactions involving readers. Interactions among the inanimate variables day, session, and theme topic were hypothesized to be secondary in importance, and all interactions of batches (to which students had been randomly assigned) with the other variables were given tertiary importance.

This design was adopted because it allows one to analyze the variance into 64 variance components, each with a single degree of freedom. With the variance thus finely analyzed it is possible, if thought desirable, to make adjustments in the mean scores leading to higher reliability of the criterion themes. We are indebted to Mr. Wiley and Professor Stanley for this footnote.

from whose data and reports it was inferred that for the purposes of the present study the rating of themes for specific factors would be too unreliable to be of much use.

In rating the themes, the readers used a 4-point scale: 4 for good, 3 for above average, 2 for below average, and 1 for bad. Scales of odd numbers were rejected because of the tendency for scores to accumulate at the convenient hypothetical mid-point. Scales with many points were rejected because reliability suffers when readers try to make their distinctions too subtle. It seemed, then, that the choice lay between the 4-point and the 6-point scales. On the basis of a pilot reading experiment (described in Appendix C) done the week before the theme reading, the 4-point scale was chosen.

On each of the evenings of June 25-28, a team of four readers met in Madison for orientation before doing their rating the next day. To help in instructing the readers, the project staff had collected 12 impromptu themes representing the four topics and written under conditions similar to those for the writing of the criterion and antecedent themes.[20] These orientation themes were considered typical of the Wisconsin High School students' impromptu writing and representative of the different points on the 4-point scale. By staff consensus, each of these themes clearly merited either a 1, 2, 3, or 4 when considered in relation to each other. They were presented to each group of readers, and the resulting discussion helped to establish standards and criteria.

---

[20] We are indebted to Mr. Leonard Schweitzer, acting head of the English Department of the New Trier Township High School, Winnetka, Illinois, and to the four New Trier teachers whose English classes wrote these themes.

The readers were instructed especially in two things. First, it was stressed that they should read their sets of themes on a relative basis, accepting standards set by the staff through the model themes. It should be made clear that readers were not asked to abandon standards, but merely to forego using any absolute or foreign standards. This meant that they had to suppress the particular standards they had grown accustomed to in their teaching situations. This procedure was necessary to ensure maximum reader reliability, and was quite proper from the fact that the treatment effects on experimental subjects were analyzed and interpreted on a relative basis.

The second thing stressed was that the readers should base their evaluations not on particular but on general impressions. That is, they were urged to react, for example, to a theme's over-all punctuation rather than to one misplaced comma, despite the personal annoyance they might find in such a point. To help them respond to the general character of a theme and avoid being misled by really atypical instances, they were encouraged to read much more rapidly than they had been used to, and to read as judges of ability reflected in an artifact of remote origin rather than as teachers personally involved and mindful of correction or reward. In orientation to criteria, the familiar factors of content, organization, and use of language were presented and discussed. All the readers seemed willing to accept these, and in that order of importance.

On each of the days of June 26-29, then, the reading team instructed the evening before read the antecedent and the criterion themes within the design described. For each set (which had several "check" themes besides the 71), the readers averaged about 90 minutes. They worked in a large air-conditioned room, isolated by screens. They recorded the single scores for a set of themes on a sheet typed with the identifying numbers in the order of the set of themes. Finally, it should be noted that the readers were ignorant of the experiment. Before their selection they had been asked how much they knew of it, and the most positive answers were that it had been vaguely heard of.

After the reading was done, the criterion themes were measured objectively. Several teaching assistants in the University's Department of English counted the following: sentence errors (fragments, run-ons, and comma faults), certain grammatical errors (faulty agreement and incorrect word forms), and misspellings. It was thought that among the feasible objective criteria, these would provide the best index of differences in the ability to use standard mechanics of writing.

One other procedure needing mention is the objective testing done in June, 1963. As part of their spring semester final examination in English, all students took an alternate form of the Iowa Test of Educational Development #7, Ability to Interpret Literary Materials, that they had taken before the experiment. The scores of the 10th, 11th, and 12th graders were used to measure the possible effects of the greater amounts of literary study given control classes in those grades during the periods the experimental classes were being instructed in composition skills.

V. ANALYSES OF THE DATA AND FINDINGS

The main criteria for testing all three hypotheses came from two period-long (56 minutes) impromptu themes written by all subjects during the last week in May. Just one subjective measure was taken of these themes. Each theme was rated by four readers for general merit on a 4-point scale. These four scores were averaged, and the mean score became a single rating for the theme. The reliability for these mean scores was estimated to be 0.83.[21] At this point, the two mean scores for each individual's criterion themes were averaged to obtain a single rating of his writing ability as measured at the end of the experiment. It was assumed that this mean of mean scores would represent an individual's writing ability better than would any single score.

These 283 scores were then analyzed to detect possible effects from the experimental variables. A three-way or three-factor analysis of variance was followed in processing the data. To the treatment factor (i. e., control vs. experimental) were added the factors of sex of student and grade level to detect possible interactions between these factors and the various treatments. Since all three factors (sex, grade, and treatment or method of instruction) were fixed effects, the fixed effects model was followed using the within cells mean square as the appropriate error term for testing all main effects and interactions. By interaction is meant the phenomena whereby differences in one variable are not consistent across the levels of the variables with which it is interacting. For example, if method A gives decidedly better results than method B at one grade level but not at another, one says that the method effects are interacting with the grade effects.

---

[21] The reliability of each set of themes (A, B, C, and D) was estimated using the formula, Reliability of *k* measurements= $1 - \frac{\text{Mean Square within}}{\text{Mean Square between}}$. These four values were averaged and their mean value taken as the estimated reliability.

The computations involved in the analyses that follow were done at the University's Numerical Analysis Laboratory. The computer program used was the "Biomed 14, General Linear Hypothesis," written by the Division of Biostatistics, UCLA. This program can analyze unbalanced analysis of variance and covariance designs according to stated hypotheses.

Experiment one was designed to test the hypothesis that a substantial increase in the amount and frequency of writing alone will result in better writing. To make the hypothesis amenable to statistical tests, the null form was assumed. That is, it was assumed that there would be no differences in the mean criterion theme scores between control and experimental groups.

Table 2 is a summary of the numbers of subjects in each of the various factor-level combinations, or "cells," in experiment one.

Table 2: Numbers of Subjects in Factor-Level Combinations for 8th and 9th Grade Experiment

Grade	Sex	Controls	Experimentals	Total by sex	Total
8	male	12	12	24	47
	female	12	11	23	
9	male	10	11	21	51
	female	15	15	30	

The controls were monthly writers and the experimentals were weekly writers.

The analysis of variance for experiment one yielded the following results. The mean score on the criterion themes for the 49 control subjects was 2.59, while the mean score for the 49 experimental subjects was 2.53. This difference is not significant. That is, a difference as large as the obtained 0.06 could easily have occurred by chance.

The mean score for the girls was 2.68 and for the boys 2.41. This difference proves significant at the .10 level. In other words, a difference this large is expected to occur by chance no oftener than one time out of ten.

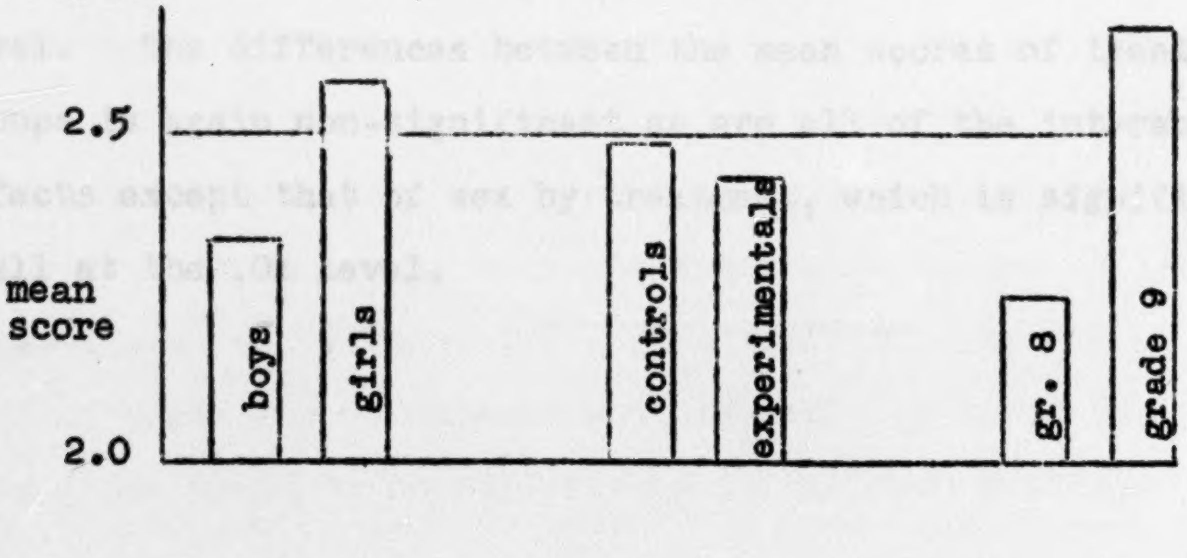
The mean score for the 8th grade subjects was 2.31 and for the 9th graders 2.79. This is significant at the .001 level (meaning that for this difference to occur by chance is highly unlikely).

All of the interaction effects proved to be non-significant--well within the limits of chance. *F* ratios and levels of significance for all main effects and interactions appear in Table 3.

Table 3: Analysis of Variance Results for Experiment One Using the Mean Score on Two Criterion Themes as the Dependent Variable

Source of Variation	*df*	*F* ratio [1, 90]	Significance
Sex	1	3.77	.10
Grade	1	17.47	.001
Treatment	1	.43	N. S.
Sex x Grade	1	2.76	N. S.
Sex x Treatment	1	.05	N. S.
Grade x Treatment	1	1.77	N. S.
Sex x Grade x Treatment	1	1.05	N. S.

The mean scores for the treatment factors are represented in this graph:



Examination of the above results raised the question whether the use of analysis of covariance applied to the data would, through increasing the power and precision of the statistical tests, bring out differences which the analysis of variance did not detect. (By increase in power is meant an increase in the probability that real differences will be detected.) An analysis of covariance was performed using five measures as covariates. These were the mean score on the first antecedent impromptu theme, the mean score on the second antecedent impromptu theme, the total score on the Nelson-Denny reading test, the SCAT verbal subtest score, and the score on the ITED #3, Correctness and Appropriateness of Expression. The sources for all of these measures had been provided during the first two weeks of the school year.

The results of the analysis of covariance appear in Table 4. These results tend to verify those obtained in the analysis of variance. The F ratio for the grade effect is 8.42, which is significant at the .01 level. The F ratio for the sex effect is 3.07, which is significant at the .10 level. The differences between the mean scores of treatment groups is again non-significant as are all of the interaction effects except that of sex by treatment, which is significantly small at the .01 level.

Table 4: Analysis of Covariance Results for Experiment One Using Mean Scores on Two Criterion Themes as the Dependent Variable. (Covariates Are Listed in Tables 11-16)

Source of Variation	df	F ratios [1, 85]	Significance
Grade	1	8.42	.01
Sex	1	3.07	.10
Treatment	1	.14	N. S.
Grade x Sex	1	1.60	N. S.
Grade x Treatment	1	.33	N. S.
Sex x Treatment	1	.00004	.01
Grade x Sex x Treatment	1	.27	N. S.

Experiment two (conducted in the upper grades) was more complicated. This experiment was designed to test two hypotheses: 1) other factors being equal, those students having a significantly greater amount of functional composition skills learning activities will develop writing ability superior to the students having the smaller amount, and 2) immediate-feedback tutoring is an effective composition skills learning activity (i.e., other factors being equal, tutored students will develop better writing ability than non-tutored students). As with the hypothesis tested in the lower-grades experiment, for statistical purposes the null form of these two hypotheses was assumed. (It was assumed that there would be no differences in the mean criterion theme scores between control and experimental students in each grade 10 through 12.)

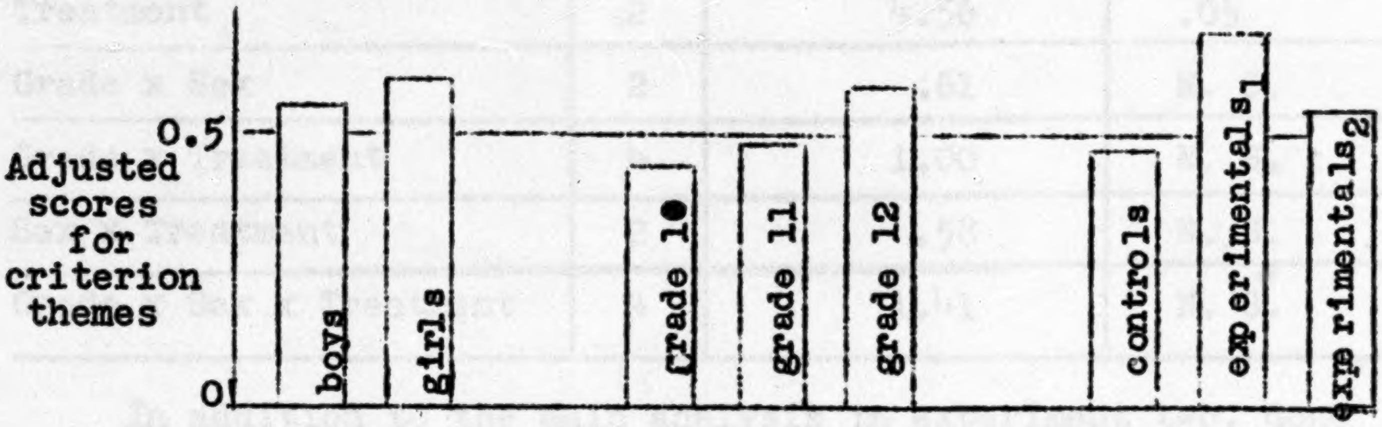
The following table is a summary of the numbers of subjects in each of the various cells in the second experiment:

Table 5: Numbers of Subjects in Factor-Level Combinations for Experiment in Grades 10, 11, and 12

Grade	Sex	Controls	Experimentals$_1$ (non-tutored)	Experimentals$_2$ (tutored)	Total
10	male	9	6	19	64
	female	10	14	6	
11	male	14	9	8	55
	female	6	8	10	
12	male	13	11	5	66
	female	7	13	17	
Totals		59	61	65	185

The analysis of covariance for experiment two yielded the following results. The mean adjusted scores of the criterion themes (adjusted through analysis of covariance) show that the differences between the three treatment groups (control, first experimental, and second experimental) are significant at the .05 level. In other words, differences of the size obtained are expected to occur by chance only one time in twenty. As for the other two main effects, grade and sex, these scores show grade differences significant at the .10 level (likely to occur by chance only one time in ten), and a sex difference in favor of the girls significant at that level.

The following graph represents the adjusted mean scores for these main effects:



The scale of these mean adjusted scores is different from the original 4-point scale because of the adjustment through analysis of covariance.

All of the interactions among factors were non-significant. This means that the treatment differences were much the same for each grade, the sex differences much the same for each grade, and the sex differences much the same within each treatment.

The main effects and interactions are represented in the following table.

Table 6: Analysis of Covariance Results for Experiment Two Using Mean Scores on Two Criterion Themes as the Dependent Variable

Source of Variation	df	*F* ratio [1, 161]	Significance
Sex	1	3.24	.10
Grade	2	3.68	.10
Treatment	2	4.56	.05
Grade x Sex	2	.61	N. S.
Grade x Treatment	4	1.00	N. S.
Sex x Treatment	2	.58	N. S.
Grade x Sex x Treatment	4	1.41	N. S.

In addition to the main analysis in experiment two, done with subjective ratings of impromptu themes, several objective measures of these themes were taken for further analysis: 1) misspellings, 2) faulty agreement and incorrect word forms, and 3) sentence errors. Analysis of covariance using these measures of the 10th, 11th, and 12th grade themes as dependent variables yielded the following significant results. Control subjects made more sentence errors than experimentals (significant at the .05 level). Boys used more incorrect word forms than did girls (.05 level). All other differences among the objective measures were insignificant.

Finally, it was thought desirable to analyze the possible effects of the experimental variables on the students' literary studies. As a single criterion the Iowa Test of

Educational Development #7, Ability to Interpret Literary Materials, was given all subjects as part of their final examination in English for the spring semester. Analysis of variance of the 8th and 9th grade scores showed that the 9th graders scored significantly better than the 8th graders (at the .05 level). But since this was not a covariance analysis, it seems highly likely that these differences existed at the beginning of the experiment as well. All other main effects and interactions were non-significant.

Table 7: Results of Analysis of Variance Using ITED #7, Ability to Interpret Literary Materials, as the Dependent Variable (Grades 8 and 9)

Source of Variation	df	*F* ratio [1, 90]	Significance
Sex	1	.79	N. S.
Grade	1	5.98	.05
Treatment	1	1.39	N. S.
Sex x Grade	1	3.97	N. S.
Sex x Treatment	1	.11	N. S.
Grade x Treatment	1	1.66	N. S.
Sex x Grade x Treatment	1	.49	N. S.

To analyze the scores of the upper graders on this test, analysis of covariance was used. There were four covariates: 1) mean scores on the two antecedent impromptu themes, 2) ITED #7 pre-test, 3) Nelson-Denny Reading Test, and 4) the SCAT verbal subtest. The only significant effect indicated by the analysis was a grade effect: the 11th graders scored better than students in the other two grades at the .05 level.

Table 8: Results of Analysis of Covariance Using ITED #7, Ability to Interpret Literary Materials, as the Dependent Variable (Grades 10, 11, and 12)

Source of Variation	df	*F* ratio [1, 161]	Significance
Sex	1	.35	N. S.
Grade	2	6.24	.05
Treatment	2	.33	N. S.
Sex x Grade	2	.60	N. S.
Sex x Treatment	2	1.73	N. S.
Grade x Treatment	4	.21	N. S.
Sex x Grade x Treatment	4	.52	N. S.

In terms of the objectives of the study, the findings can be summarized as follows. In grades 8 and 9 the experimental groups wrote about four times as much and four times as often as the control groups. Great care was taken to keep all other relevant factors equal. Under these conditions, analysis of variance revealed no difference between the two treatments (i. e., mere writing done once a week vs. mere writing done once a month) as measured by scores for general merit given by four readers to each of two criterion themes. Although these control and experimental groups had been randomly selected, an examination of certain pre-experimental measures (see Tables 11-16) for the 98 subjects indicated that the control groups had slightly more initial ability than the experimental groups. An analysis of covariance using these pre-measures (which in effect compensated for the initial differences) merely confirmed the findings of the analysis of variance, that no post-experimental difference in writing ability existed between the two groups. Neither analysis revealed any interaction of treatment with grade level. This means that the equality of writing ability between control and experimental groups was practically the same at each grade level. (In other words, the lack of measurable interaction effect between treatment and grade level indicates that no cancellation factor was operating--that is, where the control class in one grade and the experimental class in the other performed significantly better.)

In grades 10, 11 and 12, the control groups wrote one theme a month which served as the basis for about 2 and 1/2 days of functional composition instruction. The experimental$_1$ groups wrote a weekly theme through which they were given about four times as much functional instruction as the controls. The experimental$_2$ groups were given the same curriculum as the first experimentals, but were tutored with the maximum possible amount of immediate feedback during one session a week. Analysis of covariance revealed that both experimental groups showed post-experimental writing ability superior to that of the control groups. The non-tutored experimentals were better than the controls at the .05 level. The performance of the tutored experimentals fell about halfway between the other two, and is not significantly different from either. No significant interaction effects were detected. Thus, the differences between the three groups were essentially the same at each of the three grade levels.

VI. CONCLUSIONS AND IMPLICATIONS

Because the final criterion performances of 98 8th and 9th grade students showed no effects due to substantially different amounts and frequencies of writing per se, the following conclusion can be drawn. The popular belief that more writing alone means better writing is untenable, at least for the junior and senior high schools. Previous research had strongly implied that this assumption about student performance is false for the college level. Now, on the basis of the negative findings in this study, where the same degree of ineffectiveness turned up in both the 8th and 9th grades, it is reasonable to assume that neither students in the upper grades nor those in the 7th grade are benefited by a great increase in the amount and frequency of mere writing.

It might be asked, how conclusive are these findings? As for the care exercised by the staff to remove bias and contamination from the design, to generate in the classrooms a really meaningful experimental variable, to manage a competent and reliable evaluation of criterion performance, and to to analyze the data fairly and thoroughly, the reader must judge for himself from earlier parts of this report. It should be taken into account also that the findings show differences in the 8th and 9th grade criterion data which raise whatever validity is given the experiment as a whole. The analyses detected a highly significant difference between the readers' estimate of the general performance at each grade level, and a smaller and less significant difference between their opinions of the general ability of the boys and girls. The point is, because these two differences (grade and sex) and their relative size are precisely what should be expected, their detection increases the probability that the highly non-significant differences shown by the same analyses reflect accurately the real situation. Finally, the fact is that where other instructional factors are equal or absent, negative results from an increase in mere writing seem natural.

The conclusions to be drawn from experiment two are somewhat more complicated. It seems clear that when the total amount of the four functional composition learning activities is increased from about 1/8 the total instructional time in English to about 1/2, the development of writing ability is benefited to about the same significant degree in each of the senior high school grades. What inference to make from the fact that, in all three grades, the tutored groups performed about midway between the controls and the other experimentals is not certain. Either the tutoring had a negative effect tending to counteract the benefit of increasing the other four functional activities, or it had no appreciable effect and the scores for this group represent a chance downward fluctuation from the general performance of experimental subjects. No doubt different reasons to support each inference can be found. But regardless of the inference one prefers, the conclusion remains that the kind of tutoring given, in the amounts given, was in comparison to the functional activities an ineffective activity.

The contribution of this research toward distinguishing effective from ineffective composition skills learning activities at various secondary grade levels lies in the following general conclusions. The activity of writing in and of itself is fruitless. The activity of tutoring, even where immediate feedback is possible, either is negative, or ineffective, or does not increase the effectiveness of the four more ordinary and feasible functional activities, or to be effective must be given in such great amounts as to be unfeasible. Because no grade differences within treatments were detected, it seems reasonable to apply these conclusions to all grade levels 7 through 12.

What are the implications of these conclusions for English teachers and their classrooms? First, they should not assign or elicit any writing for the purpose of developing composition ability unless this writing becomes the vehicle for functional instruction. Second, they can reduce greatly the time they spend in "tutoring" writing and conferring about compositions if they give adequate functional composition instruction in the classroom.

But what is an adequate amount of these functional activities? It can be inferred in part from the facts that writing per se is ineffective and functional instruction for about half the time spent in English is significantly more effective than one-fourth that amount. Until more is known, the best inference seems to be that to teach composition effectively, teachers should give a weekly writing task on which they base about 2 and 1/2 days of practical explanation, student practice, discussion, revising, rewriting, etc.

It should be asked whether to devote this much time to teaching writing would endanger other phases of the English program. To answer this, the results of the post-experimental test "Ability to Interpret Literary Materials" should be considered, though they are not conclusive evidence. No treatment effects on this ability were detected through the tests. When it is remembered that nearly every composition written by the 10th, 11th and 12th graders was some kind of expository exercise related to the literature, the following inference might be drawn. Time lost to literary study per se is compensated for by writing tasks that are literature-centered. If this particular relationship between the literature and composition phases is established, then one-half the English program can safely be devoted to functional instruction in expository writing.

Another implication of the present study is for future research. Those who consider the findings about tutoring and mere writing essentially sound might narrow the field for classroom experimentation to the four functional activities. It might be fruitful to try to determine whether one of these functional activities is significantly more effective than the others, or whether other comparative amounts than those tested here give significantly different effects.

Finally, the implication of these conclusions for demonstration activities should be mentioned. The final difference in writing ability between control and experimental subjects in experiment two was caused by an instructional method of a certain kind. Superficially, this method seems conventional because one assumes that each of its activities has been used for a long time by an infinite number of teachers of all sorts. But it is really doubtful that many English teachers unite precisely these activities in an expository task with the single aim of affecting the function of writing ability in a meaningful situation. That this can be done effectively, even by inexperienced teachers, is proved by the present study. No doubt further experimentation is desirable, but along with it, perhaps, should be the demonstration in ordinary school settings of the effective methodology developed for experimental purposes here.

VII. APPENDICES

Appendix A. Tables Not Included in Text, Experiment One

The tables in this appendix summarize the data collected on the eighth and ninth grade subjects. Data are presented as the mean score for each sex by grade by treatment combination as well as the mean score for each level of the three factors averaged across the other two factors.

Table 9: Experiment One. Numbers of Subjects in Each of the Factor-level Combinations for the 8th and 9th Grade Experiment

Grade	Sex	Control	Experimental	Total by Sex	Total
8	Male	12	12	24	47
	Female	12	11	23	
9	Male	10	11	21	51
	Female	15	15	30	
Total by Sex	Male	22	23	45	
	Female	27	26	53	
Total		49	49		98

Control-Monthly

Experimental-Weekly

Table 10: Experiment One. Mean General Merit Scores on Criterion Themes for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	2.78	2.76	2.79
	F	2.71	2.88	
8	M	2.15	2.06	2.31
	F	2.70	2.33	
Total by Sex	M	2.43	2.40	2.41
	F	2.71	2.64	2.68
Total		2.59	2.53	2.56

Table 11: Experiment One. Mean General Merit Scores on Antecedent Theme A for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	2.80	2.25	2.53
	F	2.53	2.55	
8	M	1.88	1.96	2.09
	F	2.54	1.96	
Total by Sex	M	2.30	2.10	2.19
	F	2.54	2.30	2.42
Total		2.43	2.20	2.32

Table 12: Experiment One. Mean General Merit Scores on Antecedent Theme B for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	2.13	2.16	2.25
	F	2.28	2.37	
8	M	1.88	1.94	2.04
	F	2.29	2.05	
Total by Sex	M	1.99	2.04	2.02
	F	2.29	2.23	2.26
Total		2.15	2.14	2.15

Table 13: Experiment One. Mean Scores on SCAT-Verbal Subtest for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	297.20	296.18	292.88
	F	287.67	292.80	
8	M	290.08	283.83	287.47
	F	291.47	284.27	
Total by Sex	M	293.32	289.74	291.49
	F	289.37	289.15	289.26
Total		291.14	289.43	290.29

Table 14: Experiment One. Mean Score on Nelson-Denny Reading Test for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	76.60	71.45	67.98
	F	58.33	69.27	
8	M	62.33	45.67	59.49
	F	70.58	59.36	
Total by Sex	M	68.86	58.00	63.31
	F	63.78	65.07	64.42
Total		66.06	61.76	63.91

Table 15: Experiment One. Mean Scores on ITED, No. 3, Correctness and Appropriateness of Expression for All Factor-level Combinations

Grade	Sex	Control	Experimental	Total
9	M	68.80	71.55	67.71
	F	66.67	68.80	
8	M	57.67	54.83	60.09
	F	67.58	60.27	
Total by Sex	M	62.73	62.83	62.78
	F	67.07	65.19	66.15
Total		65.12	64.08	64.60

Table 16: Experiment One. Mean Scores on ITED, No. 7, Ability to Interpret Literary Materials, Administered June 1963

Grade	Sex	Control	Experimental	Total
9	M	56.00	55.36	54.06
	F	52.33	53.53	
8	M	46.42	42.83	48.24
	F	56.17	47.45	
Total by Sex	M	50.77	48.83	49.78
	F	54.04	50.96	52.53
Total		52.57	49.96	51.27

Appendix B. Tables Not Included in Text, Experiment Two

The tables in this appendix summarize the data collected on tenth, eleventh and twelfth grade subjects. Data are presented as the mean score for each sex by grade by treatment combination as well as the mean score for each level of the three factors averaged across the other two factors.

Table 17: Experiment Two. Numbers of Subjects in Each of the Factor-level Combinations

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	13	11	5	66
	F	7	13	17	
11	M	14	9	8	55
	F	6	8	10	
10	M	9	6	19	64
	F	10	14	6	
Total by Sex	M	36	26	32	94
	F	23	35	33	91
Total		59	61	65	185

Table 18: Experiment Two. Mean General Merit Scores Given Criterion Themes (Theme C and D) by Four Reader Teams

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	3.21	3.18	2.75	3.17
	F	3.11	3.23	3.24	
11	M	3.03	2.72	2.70	2.93
	F	2.94	3.11	3.20	
10	M	1.96	3.04	2.76	2.77
	F	2.88	3.05	2.92	
Total by Sex	M	2.83	2.99	2.74	2.84
	F	2.89	3.13	3.17	3.08
Total		2.85	3.07	2.96	2.96

Table 19: Experiment Two. Mean General Merit Scores Given Antecedent Themes (Theme A and B) by Four Reader Teams.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	3.10	2.90	2.63	3.01
	F	2.80	3.02	3.21	
11	M	3.05	2.45	2.39	2.70
	F	2.44	2.58	2.93	
10	M	1.92	2.63	2.34	2.52
	F	2.76	2.75	2.81	
Total by Sex	M	2.79	2.68	2.42	2.63
	F	2.69	2.81	3.05	2.87
Total		2.75	2.76	2.74	2.75

Table 20: Experiment Two: Mean Scores on Iowa Tests of Educational Development, No. 3, Correctness and Appropriateness of Expression for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	77.31	73.45	67.00	76.76
	F	79.86	75.85	80.76	
11	M	76.71	69.89	65.75	74.29
	F	78.17	71.25	81.80	
10	M	61.00	75.17	70.58	72.39
	F	75.20	75.36	80.83	
Total by Sex	M	73.00	72.62	68.81	71.47
	F	77.39	74.61	81.09	77.66
Total		74.71	73.75	75.04	74.51

Table 21: Experiment Two. Mean Scores on Nelson-Denny Reading Test for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	102.62	93.45	92.60	94.83
	F	85.43	89.23	98.59	
11	M	97.43	80.56	82.88	87.69
	F	77.33	84.50	93.10	
10	M	62.22	86.33	80.21	79.14
	F	74.60	83.21	92.00	
Total by Sex	M	90.50	87.35	82.81	87.01
	F	78.61	85.74	95.72	87.56
Total		85.86	86.43	89.37	87.28

Table 22: Experiment Two. Mean Scores on SCAT Verbal Subtest for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	50.00	49.55	44.40	49.52
	F	50.00	47.85	51.71	
11	M	48.93	45.33	45.13	46.84
	F	45.67	45.88	49.10	
10	M	39.44	47.33	43.00	44.05
	F	43.40	46.50	46.33	
Total by Sex	M	46.94	47.19	43.75	45.93
	F	46.00	46.86	49.94	47.76
Total		46.58	47.00	46.89	46.83

Table 23: Experiment Two. Mean Scores on SCAT Quantitative Subtest for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	41.46	42.00	37.40	38.12
	F	36.71	35.54	35.82	
11	M	42.86	37.33	39.86	38.38
	F	37.33	37.25	33.40	
10	M	38.78	40.67	39.21	38.11
	F	36.90	36.50	36.83	
Total by Sex	M	41.33	40.08	39.09	40.22
	F	36.96	36.31	35.27	36.10
Total		39.62	37.92	37.15	38.19

Table 24: Experiment Two. Mean English Grade Point Average for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	2.76	2.46	2.56	2.84
	F	2.69	2.94	3.21	
11	M	3.11	2.39	2.58	2.93
	F	3.20	2.90	3.33	
10	M	1.79	3.37	2.58	2.82
	F	2.93	3.00	3.95	
Total by Sex	M	2.65	2.65	2.58	2.63
	F	2.93	2.96	3.38	3.10
Total		2.76	2.82	2.99	2.86

Table 25: Experiment Two. Mean Scores on ITED, No. 7, Ability to Interpret Literary Materials. Administered Sept. 1962.

Grade	Sex	Control	Expt. 1	Expt. 2	Control
12	M	59.08	53.27	52.20	54.77
	F	52.29	52.00	56.35	
11	M	61.36	53.00	56.13	58.21
	F	56.17	57.88	61.70	
10	M	40.33	54.17	46.95	50.14
	F	51.10	54.64	58.83	
Total by Sex	M	55.28	53.38	50.06	52.98
	F	52.78	54.40	58.42	55.45
Total		54.31	53.97	54.31	54.19

Table 26: Experiment Two. Mean Scores on ITED, No. 7, Ability to Interpret Literary Materials. Administered June 1963.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	64.15	61.36	58.20	60.08
	F	58.43	58.00	62.76	
11	M	60.14	58.11	55.50	58.83
	F	57.00	57.00	62.90	
10	M	49.00	61.50	56.00	58.13
	F	58.40	62.00	65.67	
Total by Sex	M	58.81	60.27	56.72	58.50
	F	58.04	59.37	60.88	59.58
Total		58.51	59.75	58.83	59.03

Table 27: Experiment Two. Mean Number of Misspellings per Theme (Theme C and D) for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	3.00	3.09	2.90	3.02
	F	3.00	4.27	2.03	
11	M	3.11	2.33	6.82	3.65
	F	2.69	4.88	2.60	
10	M	10.50	2.25	4.00	4.41
	F	2.60	4.39	1.75	
Total by Sex	M	4.94	2.65	4.53	4.17
	F	2.74	4.46	2.15	3.19
Total		4.08	3.69	3.32	3.69

Table 28: Experiment two. Mean Number of Wrong Word Forms per Theme (Theme C and D) for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	.77	.46	.92	.54
	F	.57	.92	.00	
11	M	.71	1.11	1.19	1.12
	F	.50	3.19	.35	
10	M	2.56	.59	.76	1.17
	F	1.30	1.11	.92	
Total by Sex	M	1.19	.71	.89	.96
	F	.87	1.51	.27	.90
Total		1.07	1.17	.58	.93

Table 29: Experiment Two. Mean Number of Run-on Sentences, Sentence Fragments, and Comma Faults per Theme (Theme C and D) for All Factor-level Combinations.

Grade	Sex	Control	Expt. 1	Expt. 2	Total
12	M	1.12	.32	1.50	.95
	F	1.29	1.00	.91	
11	M	1.11	.56	1.38	1.11
	F	1.00	1.50	1.15	
10	M	3.33	.75	.77	1.20
	F	1.20	.82	.33	
Total by Sex	M	1.67	.50	1.02	1.12
	F	1.17	1.04	.97	1.05
Total		1.47	.81	.99	1.09

Appendix C: Pilot Experiment to Compare Four-Point and Six-Point Composition Rating Scales

Two factors had to be considered in choosing the scale to be used by readers in rating the antecedent and criterion themes. As the number of points on a scale is increased, and a rater is forced to make finer distinctions, he becomes less reliable. There is less agreement between an individual's rating one theme and the next, and less agreement from reader to reader. On the other hand, as the number of points on a scale is decreased, the power to discriminate meaningfully among themes is lost. A two-point good-bad scale may be reliable, but meaningful differences would be lost because of the coarseness of just two categories. In the judgment of the chief investigator and the design specialist, the only feasible choice lay between the 4-point and the 6-point scales. To test the relative reliability and discriminability of the two, a small reading experiment was conducted.

For this, 16 teaching assistants from the University's Department of English were chosen. They read in two teams of four during each of two sessions, alternating the 4- and 6-point scales. That is, each scale was employed by four readers during each of two sessions. For this reading, one set of 71 themes written on topic D (the general idea that anger is a dangerous feeling, etc.) was used. First, the chief investigator, his associate for supervising readers, Professor Kurtz, and project staff reached a consensus on the general merits of 13 themes from this set, and used them for orienting the readers to standards and criteria. The remaining 58 themes, a random selection stratified among the experimental subjects by grade, sex and treatment, were then read by the 16 raters in two sessions, four raters using the 4-point scale and four using the 6-point scale during each session. The chief investigator supervised the 4-point raters in session one and the 6-point in session two, and Professor Kurtz did vice versa. All 16 raters read the same 58 themes which, though in random order, were in the same order for each rater. The readers were assigned to groups randomly. It was assumed that this design controlled bias due to differences between sessions and between supervisors.

The reliabilities for the mean scores on each scale were computed using the formula 1 - [Mean Square within themes/Mean Square between themes]. From these the following reliabilities were estimated:

4-point scale: .88
6-point scale: .88

To estimate the relative discriminability or sensitivity of the two scales, it was necessary to have an independent criterion measure of the composition skill of the 58 subjects whose themes were read. Because a standardized test was felt to be best for this purpose, the Iowa Test of Educational Development #3, Correctness and Appropriateness of Expression, was used. The sets of scores on each scale were correlated with the test scores, with the following results:

4-point scale: r=.63

6-point scale: r=.64

Thus, the pilot reading experiment showed that with careful controls, no statistically significant differences exist between the 4-point and 6-point scales.

However, the two supervisors found a real difference between the scales in administering them. The readers were much more easily oriented to the 4-point scale, and they read the themes faster using the 4-point. For these reasons, the 4-point scale was used for the reading of the antecedent and criterion themes.

Appendix D: Impromptu Theme Topics and Writing Situations

1. Topic A (given September 13, 1962)

"'He who hesitates is lost,' and 'Look before you leap,' seem to be two contradictory proverbs. Write a well-organized essay of at least 300 words in which you either

Explain which of the two proverbs gives better advice, or

Discuss how the two proverbs can be reconciled."

2. Topic B (given September 19, 1962)

"Sometimes youngsters feel that older people should treat them more as adults than they generally do. Sometimes older people feel that youngsters are now considered too adult. The question seems to be: at what age do people in our society become adult? Write a well-organized essay of at least 300 words in which you discuss this question."

3. Topic C (given May 24, 1963)

"We often hear people say that 'the best things in life are free.' After you have given this idea some thought, write a well-organized essay of at least 300 words in which you either

Discuss why you feel the idea is basically true, or

Discuss why you feel the idea is basically false, or

Discuss why you feel the idea cannot be considered basically either true or false."

4. Topic D (given May 27, 1963)

"According to certain moral and religious teachings, anger is a dangerous feeling and should be carefully controlled. Consider this idea in relation to what you know human nature and right and wrong. Write a well-organized essay of at least 300 words in which you either

Discuss why you feel the idea is basically true, or
Discuss why you feel the idea is basically false, or
Discuss why you feel the idea cannot be considered basically either true or false."

All teachers urged their students to make use of the full period for the theme. They told them to revise if they wanted to, but not to waste time merely copying over a manuscript already legible. They supplied each class with two desk dictionaries. In September, they told the students on both occasions they wanted to see their best writing for purposes of diagnosis. In May, the students were told that the two impromptu themes were part of their semester examination.

Appendix E: Summary Description of Population Involved in Experiment

The Wisconsin High School students have an I. Q. range of about 90-115, with a mean of about 115. Their median family income is somewhat above that of the Madison area. About 32 percent of the fathers are of the professional class. About 25 percent are University faculty members. About 20 percent are of the managerial class. Of the others, about 11 percent are salesmen, about 7 percent artisans, and the rest (5 percent) miscellaneous.

The curricular responsibilities of the High School are entirely departmentalized, and the instructional programs are quite academic. Most of the students are serious, and it should be noted that unilaterally they seem more interested in subject matter per se, or what might be called knowledge for its own sake, than do students of similar socio-economic milieus elsewhere.

Appendix F: Wisconsin High School English Teachers Participating In Experiment

Grade 8. Mr. Theodore Lemcke, Jr.
B. A., Middlebury College
M. A. T., Johns Hopkins University
Formerly taught at the Park School, Brooklandville, Md., one semester

Grade 9. Mrs. Marjorie Elliott (fall semester)
B. A., Wheaton College (Mass.)
M. A., Cornell University
Mt. Berry School, Mt. Berry, Ga., two years; Hammondsport (N. Y.) High School, two years

Mrs. Alice Mielcarek (spring semester)
B. S., University of Wisconsin
M. S., Wisconsin
Gladstone (Mich.) High School, two years; Wisconsin High School, seven years

Grade 10. Miss Mary Henley
B. A., De Pauw University
Graduate work, Wisconsin
Previous experience only as practice teacher

Grade 11. Mrs. Janet Kurtz
B. A., Carleton College
M. A., Cornell University
Summit School, St. Paul, Minn., two years

Grade 12. Mrs. Elizabeth Dowling (fall semester
B. S., University of Illinois
M. S., Illinois
Warsaw (Ind.) High School, one year; Chili (Ind.) High School, one year; Wisconsin High School, four years

Mrs. Elliott (spring semester)

Appendix G: Composition Reading Teams

Day 1. Professor David Hiatt, Carroll College
Miss Mary Dunne, St. Norbert College
Mr. Richard Smith, Laboratory School, Oshkosh
Mrs. Gretchen Schoff, Madison West High School

Day 2. Professor Giles Daeger, Marquette University
Mrs. Angelina Lenehan, University of Oklahoma
Mr. William Parkinson, Middleton High School
Miss Margaret McGillivray, Milwaukee Washington High School

Day 3. Professor George Hartung, University of Wisconsin Extension Center, Wausau
Sister Mary Chrysostum, SSND, Mount Mary College
Mr. Hoyt Rawlings, West Allis Central High School
Miss Marilyn Radke, Milwaukee Nicolet High School

Day 4. Professor Edward Laun, Wisconsin State College, Whitewater
Miss Marion Hawkins, Wisconsin State College, River Falls
Mr. Keith Livens, Monroe High School
Miss Edythe Daniel, Doudna Campus School, Platteville

BIBLIOGRAPHY

A. Works Cited in Statement of Problem

Dusel, William. "Determining an Efficient Teaching Load in English," *Illinois English Bulletin*, October, 1955, pp. 1-19.

B. Works Cited in Review and Discussion of Related Research

Bostrom, Robert, and others. "Grades as Reinforcing Contingencies and Attitude Change," *Journal of Educational Psychology*, April, 1961, pp. 112-15.

Buxton, Earl. *An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Students' Skill in Written Expression*, Stanford University, 1958 (unpublished dissertation).

Dressel, Paul, and others. "Effect of Writing Frequency upon Essay-Type Writing Proficiency at the College Level," *Journal of Educational Research*, December, 1952, pp. 285-93.

Heys, Frank. "The Theme a Week Assumption," *English Journal*, May, 1962, pp. 320-22.

Lokke, Virgil, and George Wykoff. "Doubling Writing in Freshman Composition--an Experiment," *School and Society*, December 18, 1948, pp. 437-39.

Lyman, Rollo. *Summary of Investigations Relating to Grammar, Language and Composition*. University of Chicago Press, 1929 (Supplement, 1931).

Maize, Ray. "Writing Laboratory for Retarded Students," *College English*, October, 1954, pp. 44-48.

Page, Ellis. "Teacher Comments and Student Performance: a Seventy-four Classroom Experiment in School Motivation," *Journal of Educational Psychology*, August, 1958, pp. 173-81.

Pooley, Robert. Teaching English Grammar. Appleton-Century-Crofts, 1957.

Schrom, Ned. Missouri High School English Offerings Reported by University Freshmen in Relation to Their Freshman English Placement Test Scores, University of Missouri, 1953 (unpublished dissertation).

Smith, Dora. "English Grammar Again!," English Journal, October, 1938, pp. 643-49.

Strom, Ingrid. "Research in Grammar and Usage and Its Implications for Teaching Writing," Indiana University School of Education Bulletin, 1960.

C. Works Cited in Discussion of Design and Procedures

Box, George, and J. S. Hunter. "The 2^{k-p} Fractional Factorial Designs," Mathematics Research Center Summary Report No. 218. University of Wisconsin and the United States Army, 1961.

Conant, James B. The American High School Today. McGraw-Hill, 1959.

Rand Corporation. A Million Random Digits with 100,000 Normal Deviates. Glencoe Free Press, 1955.

Remondino, C. "A Factorial Analysis of the Evaluation of Scholastic Compositions in the Mother Tongue," British Journal of Educational Psychology, November, 1959, pp. 242-51.

D. Works Used in Analyses of Data

"BIMD 14: General Linear Hypothesis," Division of Biostatistics, School of Medicine, UCLA, June 1, 1960.

McNemar, Quinn. Psychological Statistics (2nd ed.). John Wiley and Sons, 1955.

Winer, B. J. Statistical Principles in Experimental Design. McGraw-Hill, 1962.